DeFake StyleGAN Deepfake Detector for Facial Images

Piyush Singh, Suhana Sabir Khan, Sanjana Srinivas, Rohini B R
Global Academy of Technology,
Bengaluru, India
{piyush.syngh79, suhanasaxbr, sanjana.srinivas1154, rohini.br}@gmail.com

Abstract: **This research introduces DeFake, a system designed to detect highly realistic facial images generated by advanced AI techniques like StyleGAN. As AI-generated images become increasingly convincing, identifying fakes has become a vital challenge. DeFake employs deep learning models that analyze subtle patterns and artifacts left behind during the image generation process. By examining image details and hidden frequency patterns, our system can accurately distinguish between real images and AI-created fakes, even when they look almost identical to the human eye. This technology is especially useful for verifying digital content, supporting law enforcement, and enhancing cybersecurity efforts. Our approach not only detects fake images but also helps trace their origin, addressing the critical need to protect the integrity of visual media in the digital age. As AI technologies evolve, tools like DeFake are essential to maintaining trust and authenticity in digital content.**

Keywords — Deepfake detection, Generative adversarial networks, StyleGAN, Convolutional neural networks, GAN fingerprints, Frequency domain analysis

I. INTRODUCTION

The pace of advancements in Generative Adversarial Networks (GANs), especially in the latest models like StyleGAN2 and StyleGAN3, has considerably improved the authenticity of artificially generated facial images using AI. The resulting deepfake images are very hard to differentiate from actual images when viewed with naked eyes, which makes these images pose a huge threat to digital trust, authenticity, and information security online. The potential misuse of these images in misinformation campaigns, identity fraud, and other social manipulations calls for an urgent need for better forensic solutions for their detection.

Traditional deepfake verification relies extensively on traditional means of verifying through careful review or simple rule-based systems that verify for apparent flaws such as unreal facial movements and lighting reflections. But as the model gets better, these apparent flaws become less significant or even become absent altogether. This affects the efficiency of traditional verification systems and is a problem that can only be filled by the development of specialized machine learning-based forensic systems that can detect even the embedded unseen patterns.

In order to counter this issue, the proposal for the solution is the use of the DeFake deep learning forensic solution for deepfake detection and attribution. This solution will be developed utilizing the dual-branch architecture concept with the spatial and frequency domain analysis techniques. Convolutional Neural Networks will be used for the purpose of identifying the texture inconsistencies locally, and the other will be the Fast Fourier Transform technique that will be used for the purpose of analyzing the high-frequency anomalies caused by the GAN upsampling and the convolution processes.

One of the major advantages of DeFake is the fact that it not only has the capability to detect deepfake images, but it also enables source image attribution. It is capable of identifying the source track and matching the fingerprint patterns that help attribute a particular fake image to its source generator, StyleGAN2 or StyleGAN3. This makes the tool much more useful for its applications in legal investigations, journalism, and digital forensic analysis, among others, which not only demands detection but also source identification.

Further ensuring robustness and transparency is the employment of strong dataset handling techniques like data augmentation and adversarial training, which help to generalize well in a real-world setup that might have noise, compression, and size variations. In addition to that, there is also the use of Explainable AI techniques like Grad-CAM to highlight areas that inform the networks' decision-making. This enhances transparency and helps to show that it is not relying on facial identification features.

DeFake offers a promising solution to this pressing issue. With its fingerprint analysis, deep learning, and explanation capabilities, DeFake ensures correct, understandable, and efficient verification of digital content in real-time. With advancements in generative techniques, DeFake serves as a good starting point or platform for future development, such as enhancing deepfake detection techniques for videos or diffusion-based generative models.

II. LITERATURE REVIEW

[1] Yu et al. offered the first comprehensive work on GAN fingerprinting for image attribution, tackling the problem of tracing the original source of AI-generated images. This research demonstrated that every unique GAN leaves a unique, stable fingerprint on the generated image, which varies based on the GAN structure, dataset, and initialization. This technique employs a CNN-based attribution network to discriminate between real images and AI-generated images, and trace the original source of the generated image.

[2] Barni et al. (2020) developed a CNN-based system to identify GAN-generated face images using inconsistencies between color spectral bands. The model, Cross-CoNet, was able to significantly distinguish between real and GAN-generated faces. Experiments showed almost perfect results in detecting images generated b echnique

also proved robust to post-processing techniques such as blurring, rotation, and addition of noise.

[3] Zhang et al. (2019) examined detecting GAN-generated images through the simulation of unique visual artifacts that remain within images generated by GAN. The paper targeted upsampling inconsistencies as well as color and texture distributions. The paper proposed a CNN-based classifier capable of generalizing across different GAN systems.

[4] Wang et al. (2019) described an in-depth analysis on the detection of CNN-generated images, showing the existence of characteristic statistical abnormalities in early GAN-generated images. This paper proposed a universal CNN classifier that could identify real and fake images for a variety of architectures, resulting in an extremely high degree of generalization for images from novel generators.

[5] The authors designed a two-stage approach that enhances the detection of fake images produced by GANs through the sequential process of feature extraction and classification. Experiments verified that the approach outperformed traditional two-stage detectors under image compression and noise addition, emphasizing the benefits of multi-level feature fusion.

[6] Marra et al. (2019) explored if GAN-produced images retain a unique artificial trace resembling noise patterns in real cameras. The researchers proved that every GAN produces a set of unique, stable, model-dependent marks that can be used to identify the origins of the produced image, pioneering artificial trace-based deepfake detection.

[7] Neves et al. proposed GANprintR, an autoencoder-inspired method for removing GAN fingerprints from generated face images. They tested how efficiently state-of-the-art deepfake detection tools could be fooled, revealing substantial increases in detection error rates. They also provided the iFakeFaceDB benchmark dataset.

[8] Abdullah et al. (2024) performed a thorough analysis of deepfake image detection methods in dynamic adversarial settings. They tested eight existing methods and reported that none generalize well against tailored generative methods and adversarial attacks generated via foundation models such as CLIP and ViT.

[9] A comparative assessment of CNN models and Vision Transformer (ViT) models for identifying GAN-generated deepfake images was performed. The findings showed that CNNs performed better for local feature extraction, while ViTs showed better generalization for unseen GANs, emphasizing a drift towards transformer models.

[10] Guarnera et al. introduced a new concept for deepfake detection using the identification of convolutional traces — forensic patterns resulting from the generative process in GANs. Their technique employed an Expectation-Maximization (EM) algorithm, with subsequent machine-learning classifiers (SVM, LDA, KNN) achieving high accuracy across StyleGAN and StarGAN.

[11] This paper proposes the use of a locality-aware autoencoder with incremental learning to boost the generalization capability of deepfake detectors, demonstrating better cross-dataset performance on FF++ and Celeb-DF benchmarks than traditional CNN-based detectors.

[12] An explainable deepfake detection system integrates frequency domain feature analysis with an attention mechanism for improved interpretability. The system demonstrated enhanced resilience to compression, noise, and new generation attacks when evaluated with StyleGAN and ProGAN datasets.

[13] The authors proposed a self-supervised learning framework to detect AI-generated faces via exploiting stylistic inconsistencies of the W+ latent space of StyleGAN, reporting more than 93% detection accuracy over GAN and diffusion models with surpassing results on previous benchmarks.

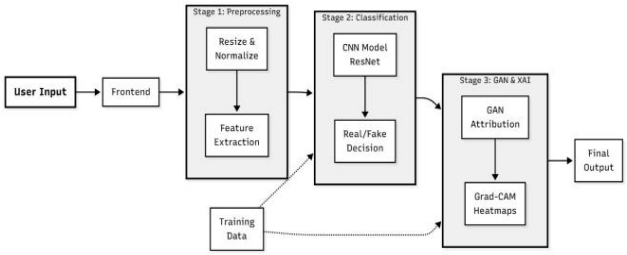

Fig. 1. Simple flow of the DeepFake detection system using CNN and Grad-CAM for clear results.

III. METHODOLOGY

The approach consists of four main stages: data processing, extraction of features, modeling, and evaluation. We will code this approach by making use of Python as well as scientific libraries.

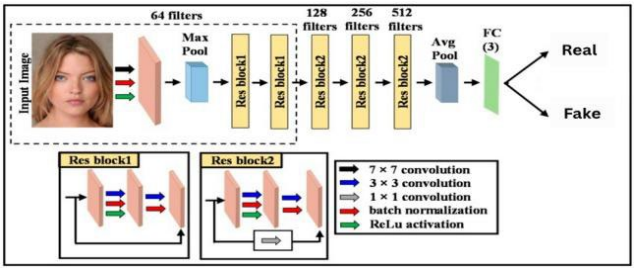

Fig. 2. CNN-based DeepFake detection architecture using residual blocks for real and fake image classification.

A. Data Acquisition

The training and testing required for developing a deepfake detection system is obtained from several publicly available benchmark sources possessing both real and synthetic facial images. Specifically, the real face images are sourced from FFHQ (Flickr-Faces-HQ, 70,000 high-resolution images) and CelebA-HQ (30,000 celebrity face images). Synthetic images are drawn from StyleGAN2 and StyleGAN3 generated outputs, as well as the FaceForensics++ (FF++) dataset which includes manipulated face videos and frames. These datasets encompass a vast range of facial characteristics, with images exposed to varied real-world conditions such as image compression, noise injection, image resizing, and other transformations.

B. Data Preprocessing

Prior to training a model on these data collections, there is a data preprocess stage aimed at improving quality. Any image

that is corrupted, incomplete or of low quality is removed. The size of all images is normalized to 256x256 pixels. In order to further increase diversity and decrease the risk of overfitting, images could be subjected to various data augmentation methods like horizontal flipping, rotation, scaling, and pixel shifting. Images have been labeled for binary classification as real or fake and multi-class classification for GAN sources.

C. Feature Extraction

The proposed system combines a two-stream approach to feature extraction to leverage both the spatial and frequency domains for artifacts extraction. The spatial approach involves the use of convolutional neural networks (CNNs), which extract RGB-based features that delineate discrepancies in textures and structural artifacts within facial images. Simultaneously, the frequency approach involves Fast Fourier Transform (FFT), which extracts spectral features to depict periodic patterns within facial images introduced by GAN synthesis in the higher frequency domains. The two-dimensional FFT is given by Equation (1), and the Power Spectral Density (PSD) derived from it is given by Equation (2).

$$F(u, v) = \Sigma\Sigma\, f(x, y) \cdot e^{(-j2\pi(ux/M + vy/N))} \quad (1)$$

F(u,v) = 2D Fast Fourier Transform, where f(x,y) is pixel intensity, M and N are image dimensions

$$PSD(u, v) = |F(u, v)|^2 \quad (2)$$

Power Spectral Density (PSD) derived from FFT

D. Model Training and Optimization

Several deep architectures are considered within the training process, with ResNet-based CNN architectures chosen for spatial feature extraction because of their stability. The use of shallow CNN architectures for frequency domain feature processing considers computational complexity. The two-stream combined network is trained end-to-end using a binary cross-entropy loss (Equation 3) for real/fake classification and a multi-class cross-entropy loss (Equation 4) for GAN source attribution.

$$L_{nCE} = -1/N\, \Sigma[y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)] \quad (3)$$

Binary Cross-Entropy Loss for real/fake classification

$$L_{CE} = -1/N\, \Sigma\Sigma\, y_{i,k} \log(\hat{y}_{i,k}) \quad (4)$$

Multi-Class Cross-Entropy Loss for GAN source attribution over C classes

E. Attribution and Reverse Fingerprinting

To make the system traceable in the forensic arena, the system uses a reverse fingerprinting technique that traces synthetic images to GAN models used to create them. Noise residuals are derived from images by subtracting a denoised version from the original, as defined in Equation (5), with GAN-specific fingerprint patterns being derived from such residuals. These fingerprints are then matched with pre-assembled fingerprints in a database using the cosine similarity measure (Equation 6), which determines similarity between patterns irrespective of scale differences.

$$R(x, y) = I(x, y) - \hat{I}(x, y) \quad (5)$$

Noise Residual R(x,y): I(x,y) is the original image, $\hat{I}$(x,y) is the denoised image

$$sim(F_1, F_2) = (F_1 \cdot F_2) / (\|F_1\| \cdot \|F_2\|) \quad (6)$$

Cosine Similarity between fingerprint vectors F_1 and F_2 for GAN attribution

Fig. 3. EfficientNetB0 model with MBConv blocks for feature extraction.

F. Evaluation and Explainability

Accuracy, precision, recall, F1-score, and the ROC-AUC score measured on several test datasets will be used as standard performance metrics for the proposed system, formally defined in Equations (7) and (8). For better interpretability and trust, Grad-CAM analysis (Equation 9) can be performed to localize image regions that contribute most strongly to the system's detections. Additionally, the system is assessed on a range of public datasets aggregated from diverse GAN sources.

$$Precision = TP / (TP + FP), \quad Recall = TP / (TP + FN) \quad (7)$$

Precision and Recall; TP=True Positives, FP=False Positives, FN=False Negatives

$$F_1 = 2 \cdot (Precision \times Recall) / (Precision + Recall) \quad (8)$$

F_1-Score: harmonic mean of Precision and Recall

$$L^c = ReLU(\Sigma_k\, \alpha^c_k\, A^k), \quad \alpha^c_k = (1/Z)\Sigma_i\, \Sigma_j\, (\partial y^c / \partial A^k_{ij}) \quad (9)$$

Grad-CAM: L^c = class activation map; α^c_k = gradient weight for feature map A^k

G. Deployment and Real-Time Testing

A demonstration interface for this interactive model has been established for easy engagement and evaluation of deepfake detection. The deployment architecture of this model is based on the API model architecture, which can be easily integrated with other applications to deploy the model. Performance parameters such as time of inference, memory consumption, and computational cost of the model have been taken into consideration to ensure that the model is scalable.

IV. RESULTS AND DISCUSSION

DeFake was evaluated over a balanced dataset of real facial images and generated images using some state-of-the-art architectures like StyleGAN2 and StyleGAN3. This balanced setting prevented class-specific bias and allowed a fair assessment of detection performance. The diversity of the dataset in terms of variations in facial attributes and generation settings helped to establish the system's generalization over different forms of synthetic content.

The proposed model achieved an overall accuracy of more than 94%, while precision, recall, and F1-score were always greater than 90%. These results show the strong performance of the system in distinguishing between real and deepfake images while keeping the false alarm rate very low. High values for recall suggest a low rate of missed detection of fake images, which is very crucial in security-sensitive applications such as misinformation detection and digital forensics. Table I presents a quantitative comparison of DeFake against established baseline models.

Fig. 4. Confusion matrix showing model performance for real and fake image classification.

This performance largely depends on the hybrid feature extraction that integrates spatial features learned through CNNs with frequency-domain features from FFT. The dual-domain approach will let the model detect both the observable texture irregularities and subtle spectral deviations introduced in the process of GAN-based image synthesis. Linked to GAN, these fingerprints are usually imperceptible to human observers; thus, forensic analysis by means of automated techniques has certain advantages compared to manual examination.

Beyond hybrid feature extraction, noise residual analysis further enhances the system's sensitivity to synthetic artifacts. This involves subtracting a denoised version of the image from the original so that the model can isolate structured noise patterns characteristic of GAN generation. The process tends to suppress semantic facial information while amplifying generator-specific traces, thereby enhancing robustness against visually convincing deepfakes.

Robustness testing by adding noise, compressing images, or using adversarial perturbations was done to assess the reliability under real-world conditions. Despite minor performance degradation under substantial distortions, overall detection accuracy remained stable, proving that DeFake is resistant to common post-processing operations. This would be important for practical deployment, as many images are compressed or otherwise altered before analysis.

Model interpretability was done through Grad-CAM visualization, which generates heatmaps showing the image regions that most influence the model's predictions. These visualizations reflect that the system focuses on artifact-prone areas like texture-rich regions and zones with frequency inconsistencies rather than on facial identity features. This instills trust in the system by showing that decisions are based on forensic evidence rather than spurious correlations.

In summary, reverse fingerprinting of the system allowed it to attribute deepfake images to their originating GAN models using residual noise matching and cosine similarity. This attribution ability has significant forensic value in terms of traceability of synthetic media sources. The model was then deployed using a user-friendly graphical interface and was further assessed for real-time inference feasibility in media verification applications, cybersecurity, and digital investigations.

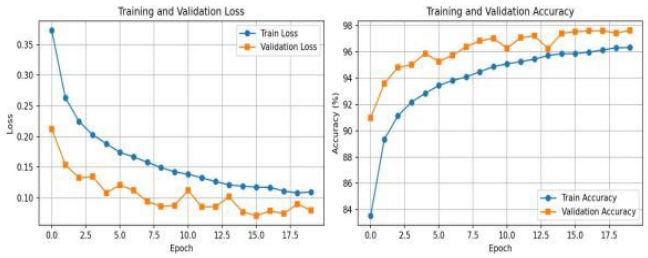

Fig. 5. Training and validation loss and accuracy curves over epochs.

These results position DeFake as a high-performance deepfake detection framework that effectively integrates accuracy, robustness, and interpretability. The system is able to use a combination of spatial- and frequency-domain features to detect GAN-specific artifacts too subtle for human observers to recognize, thus ensuring reliable detection even for very realistic-looking synthetic images.

Furthermore, explainable AI techniques and reverse fingerprinting make the system more transparent and increase its forensic credibility. Visualization of decision-making regions and the ability to attribute fake images to source GAN models make DeFake suitable for critical applications such as media verification, cybersecurity, and legal investigations.

V. CONCLUSION

DeFake is an advanced deep-learning forensic system developed to detect highly plausible deepfake facial images that are generated by the state-of-the-art generative adversarial networks such as StyleGAN2 and StyleGAN3. These fake images are made to look like real photographs and are threatening digital trust, privacy, and media integrity. This paper introduces a new dual-branch architecture in the system, which combines convolutional neural networks for spatial feature analysis with FFT-based models to capture faint GAN-derived fingerprints that are invisible to human observers. Apart from forgery detection, DeFake performs the attribution of images to specific GAN models, enhancing forensic reliability and aiding law enforcement investigations.

DeFake is trained on various datasets containing noise, compression artifacts, and adversarial perturbation with very high accuracy and robustness for most real-world conditions. It also includes explainable AI techniques such as Grad-CAM to visualize the important regions responsible for decision-making to enhance transparency and trust. This scalable and efficient system has been optimized for platforms with limited computational resources to support on-platform processing.

To tackle these emerging challenges, including deepfakes manipulated adversarially and rapidly evolving generative technologies, DeFake uses continuous model updates. It supports forensic analysis of multi-modal media and lays a basic foundation for future extension to video and audio deepfake detection. From media verification to cybersecurity and public safety, this is a comprehensive, adaptive, and dependable tool in safeguarding authenticity and countering the ever-evolving threat landscape of deepfakes.

REFERENCES

[1] J. Choi, T. Kim, Y. Jeong, S. Baek, and J. Choi, "Exploiting style latent flows for generalizing deepfake video detection," arXiv:2403.06592.

[2] T. Anand, S. Manivannan, S. Sankar, and K. Mitra, "Learning self-supervised style representations for detecting AI generated faces," in 1st Workshop on GenAI Watermarking, ICLR 2025.

[3] "An analysis of recent advances in deepfake image detection in an evolving threat landscape," in 2024 IEEE Symposium on Security and Privacy (SP), 2024. doi: 10.1109/SP54263.2024.00182.

[4] L. Guarnera, O. Giudice, and S. Battiato, "DeepFake detection by analyzing convolutional traces," in CVPR Workshops, 2020.

[5] J. Choi, T. Kim, Y. Jeong, S. Baek et al., "Exploiting style latent flows for generalizing deepfake video detection," 2021.

[6] V. N. Convertini, D. Impedovo, U. Lopez, G. Pirlo, and G. Sterlicchio, "Discrete fourier transform in unmasking deepfake images: A comparative study of StyleGAN creations," Information, vol. 15, no. 11, p. 711, 2024. doi: 10.3390/info15110711.

[7] L. Guarnera, O. Giudice, and S. Battiato, "Mastering deepfake detection: A cutting-edge approach to distinguish GAN and diffusion-model images," ACM Trans. Multimedia Comput. Commun. Appl., vol. 20, no. 11, pp. 1-24, 2024. doi: 10.1145/3652027.

[8] N. Yu, L. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," arXiv:1811.08180, 2019.

[9] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," IEEE Commun. Surveys Tuts., vol. 21, no. 3, pp. 2224-2287, 2019. doi: 10.1109/comst.2019.2904897.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in CVPR, 2022, pp. 10674-10685.

[11] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in GAN fake images," in IEEE WIFS, 2019, pp. 1-6.

[12] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in CVPR, 2020, pp. 8107-8116.

[13] Y. Lan, X. Meng, S. Yang, C. C. Loy, and B. Dai, "E3DGE: Self-supervised geometry-aware encoder for style-based 3D GAN inversion," in CVPR, 2023.